Earth Observation Image Semantic Bias: A Collaborative User Annotation Approach

Ambar Murillo Montes de Oca, Reza Bahmanyar, Nicolae Nistor, and Mihai Datcu *Fellow, IEEE*

***Abstract*—Correctly annotated image datasets are important for developing and validating image mining methods. However, there is some doubt regarding the generalizability of the models trained and validated on available datasets. This is due to dataset biases, which occur when the same semantic label is used in different ways across datasets, and/or when identical object categories are labeled differently across datasets. In this article, we demonstrate the existence of dataset biases with a sample of 8 remote sensing image datasets, first showing they are readily discriminable from a feature perspective, and then demonstrating that a model trained on one dataset is not always valid on others. Past approaches to reducing dataset biases have relied on crowdsourcing, however this is not always an option (e.g., due to public-accessibility restrictions of images), raising the question: How to structure annotation tasks to efficiently and accurately annotate images with a limited number of non-expert annotators? We propose a collaborative annotation methodology, conducting image annotation experiments where users are placed in either a collaborative or individual condition, and we analyze their annotation performance. Results show the collaborators produce more thorough, precise annotations, requiring less time than the individuals. Collaborators labels show less variance around the consensus point, meaning their assigned labels are more predictable and likely to be generally accepted by other users. Therefore, collaborative image annotation is a promising annotation methodology for creating reliable datasets with a reduced number of non-expert annotators. This in turn has implications for the creation of less biased image datasets.**

***Index Terms*—Dataset Biases, Remote Sensing Images, Semantic Image Annotation, Sensory and Semantic Gaps, User Evaluation**

I. Introduction

THE ever-growing volume and diversity of image databases increases the importance of diverse and generalizable datasets for the training and validation of content-based image retrieval methods. This raises the question: *How generalizable are the different existing datasets as training data?* Torralba and Efros [1] assessed the biases of different photographic multimedia image datasets by training a model on one dataset and testing it on another one. The authors found a low degree of generalizability across the datasets, and even expressed doubts whether existing datasets are too limited to be considered as a reflection of the real world.

Dataset biases occur when the same semantic label is used in different ways across various datasets, and/or identical object categories are labeled differently across datasets. This could be due to the dataset creation and image selection methods, which are also influenced by the purpose of the dataset. Biases can also be introduced by the annotators, due to the existing sensory and semantic gaps. The sensory gap refers to the difference between object perception with the naked eye, and the perception of the object based on the images created from sensor-recorded signals [2], [3]. The semantic gap is defined as the difference between the user and computer understanding of objects in an image [3], [4], [5], as well as the differences between various users' image understanding [6]. For remote sensing images, these biases are more pronounced due to their specific characteristics, such as resolution, perspective, and the variety of sensors being used (e.g., Synthetic Aperture Radar (SAR) or optical multispectral images), which record signals very differently from the human visual system [7].

In this article, we evaluate the biases and generalizability of eight remote sensing image datasets. These datasets were created by different authors following specific criteria (label- vs. image-based), using image products with diverse properties acquired from SAR and multispectral (optical) sensors. We train a classifier on one of the eight datasets and test it on the others from the same sensor type. The results show a significant decrease in classification performance when compared with training and testing on the same dataset. This decrease is even larger when we train on a dataset from one sensor type and test it on datasets from another sensor type. Therefore, training or validating a content-based image retrieval method on one dataset does not necessarily mean that it will generally perform well. Although this is a problem several researchers may assume exists, measuring and minimizing its effects has not been sufficiently addressed. This leads to the question: *How to create image datasets which are less biased and more representative of the real world?*

Torralba and Efros [1] recommend annotating images through crowdsourcing, which is an approach based on distributing a task across many people and integrating the individual efforts to achieve the final results, thereby reducing the effects of subjective biases. Crowdsourcing was found to be a promising procedure for creating datasets [8], and is used for image annotation through various platforms and approaches such as LabelMe [9], Amazon Mechanical Turk [8], and "Games with a Purpose" [10], [11]. Some examples using

A. Murillo Montes de Oca[+], R. Bahmanyar[+,*], and M. Datcu[*] are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany (e-mail: ambar.murillomontesdeoca@dlr.de; reza.bahmanyar@dlr.de; mihai.datcu@dlr.de).

N. Nistor and A. Murillo Montes de Oca are affiliated with the Faculty of Psychology and Educational Sciences, Ludwig-Maximilians-Universitaet, Munich, Germany (e-mail: nic.nistor@uni-muenchen.de). N.Nistor is also affiliated with Walden University, Minnesota, USA. .

[+]The corresponding authors.

[*]The authors are also affiliated with the Munich Aerospace Faculty, Munich, Germany.

publicly available satellite imagery include Tomnod[1], a crowdsourcing platform on which volunteers annotate satellite data as part of campaigns searching for specific objects. Geo-Wiki[2] is another platform using crowdsourcing for several tasks, including the improvement of land cover maps.

Crowdsourcing image annotations, however, is not always an option for remote sensing images, due to restrictions on the public access of the images, the complexities of certain images (e.g., TerraSAR-X image products [12]), or to the small number of experts available for the annotation of specific types of images. Therefore, remote sensing image datasets built via crowdsourcing are not plentiful, and expert annotations are usually relied on for dataset creation. However, due to the presence of the sensory and semantic gaps, even expert annotators can bias a reference dataset through errors in their annotations [13], or their subjective understandings [6]. This method of dataset creation is further limited by the relatively small number of experts available, and the large effort necessary to produce large scale datasets.

Although experts are relied on for image annotation, See *et al.* [14] showed that non-experts can produce annotations of the same quality as experts, depending on how the image annotation task is structured. Hutt *et al.* [15] further demonstrate that annotation task structure can have a great influence on the results. The authors found that ranking tasks (i.e., ordering images according to a specific attribute) produced the highest accuracy, inter-annotator agreement, and reliability— compared with scoring or classification tasks. This highlights the fact that task structure should be considered as an important component in designing efficient annotation methodologies.

In this article, we further study the task structure and its influence on user behavior, performance, and image understanding in annotation tasks. This will lead us to design better annotation task structures, using a reduced number of annotators, achieving more reliable image annotations. This in turn has implications for the creation of less biased image datasets. More specifically, we analyze user behavior and image understanding in individual vs. collaborative image annotation scenarios. We conduct experiments where users annotate satellite images — either collaboratively (with a partner) or individually — and we study the effects of this task structure on the efficiency and learnability of the annotation task, as well as on the quality and consistency of the annotations produced. These experiments were followed by user interviews, which assessed users' perceptions of the task, such as their confidence in the correctness of their annotations.

Our results indicate that the annotation obtained through a collaborative approach, involving 6 pairs of users, has 6.5% higher precision and 22% fewer missed labels compared with the individual users, even though collaborators spent an average of 7% less time performing the task. All users learn to identify the objects under varying conditions over time; however, the collaborators demonstrate an increased ability to identify the more difficult object categories. Furthermore, the collaborators' overall performance is considerably less variable (more than 10 times less variable), and therefore more predictable, than the individuals' performance. According to the results, the agreement between user labels among the collaborators is 79.6% while among the individuals it is 65.2%. This reflects the fact that verbal discussions on the label meanings resulted in more consistent and detailed annotations. Additionally, collaborators reported a higher interest in continuing the annotation task. This consistency and reduced variability in performance allows a consensus in annotation results to be reached with fewer users. Therefore, collaborative approaches should be further considered in designing efficient annotation methodologies, particularly when the number of available annotators is limited (e.g., in the case of non-public remote sensing images) in order to efficiently use the power of consensus.

The rest of this paper is organized as follows: Section II discusses the sensory and semantic biases existing in the eight selected remote sensing image datasets. Section III analyzes the degree of shared features between these datasets, while Section IV evaluates the generalizablility of the different datasets as training datasets. Section V discusses the effects of the collaborative task structure on user performance, and outlines the implications for creating less biased image datasets. Conclusions are presented in Section VI.

II. Sensory and Semantic Dataset Biases

Although efforts are placed to create unbiased and generalizable datasets which represent the real world, strong biases are still present [1], [6]. In this section, we experimentally demonstrate the sensory and semantic biases existing in remote sensing image patch datasets, and discuss some potential causes of these biases. Image patch datasets are usually created for patch-based analyses of high resolution remote sensing images, by cutting the full-sized scenes into smaller tiles known as "patches."

For these experiments we chose eight different datasets, selected for the diversity of their sensors, as well as for the diverse methods used to generate them. Please refer to Table I for a description of the datasets, and to Fig. 1 for sample image patches of each dataset for the categories "Urban/Residential areas" and "Agricultural fields." Datasets D1-D6 all correspond to high resolution SAR images. D1-D4 were created by the same annotator, providing the interesting possibility of comparing datasets while holding the annotator constant. The datasets were generated using an *"image-based"* approach. In this approach, a remote sensing scene is selected, and then split into image patches of a certain dimension. Since this is done disregarding image content, the resulting image patches are not centered on an object, and therefore usually their main concept is not clear. Consequently, different annotators will have different interpretations of the image patch content, caused by the existing sensory and semantic gaps, and this introduces biases into the datasets. In this dataset generation method, the number of image patches per class are not equally distributed, because categories are generated organically, meaning that they are selected to reflect existing

[1]http://www.tomnod.com
[2]http://www.geo-wiki.org

ID	Datasets	Sensor	Resolution (m)	Dataset Size (No. of Patches)	No. of Classes	No. Patches per Class	Patch Size (pixels)	Generation Method
D1	VHR SAR RU [16]	TerraSAR-X	1.25	7187	40	Not Equal	160×160	Image-based
D2	VHR SAR DE & CH [16]	TerraSAR-X	1.25	7176	42	Not Equal	160×160	Image-based
D3	VHR SAR CN [16]	TerraSAR-X	1.25	1014	7	Not Equal	160×160	Image-based
D4	VHR SAR CU [16]	TerraSAR-X	1.25	1008	9	Not Equal	160×160	Image-based
D5	15 Class TerraSAR-X [17]	TerraSAR-X	1.20	3434	15	Not Equal	160×160	Both
D6	11 Class TerraSAR-X [18]	TerraSAR-X	1.00	1100	11	Equal	160×160	Label-based
D7	UC Merced Land Use [19]	Optical (RGB)	0.30	2100	21	Equal	256×256	Label-based
D8	RS Dataset [20]	Optical (RGB)	0.50	600	12	Equal	600×600	Label-based

TABLE I
CHARACTERISTICS OF THE DATASETS

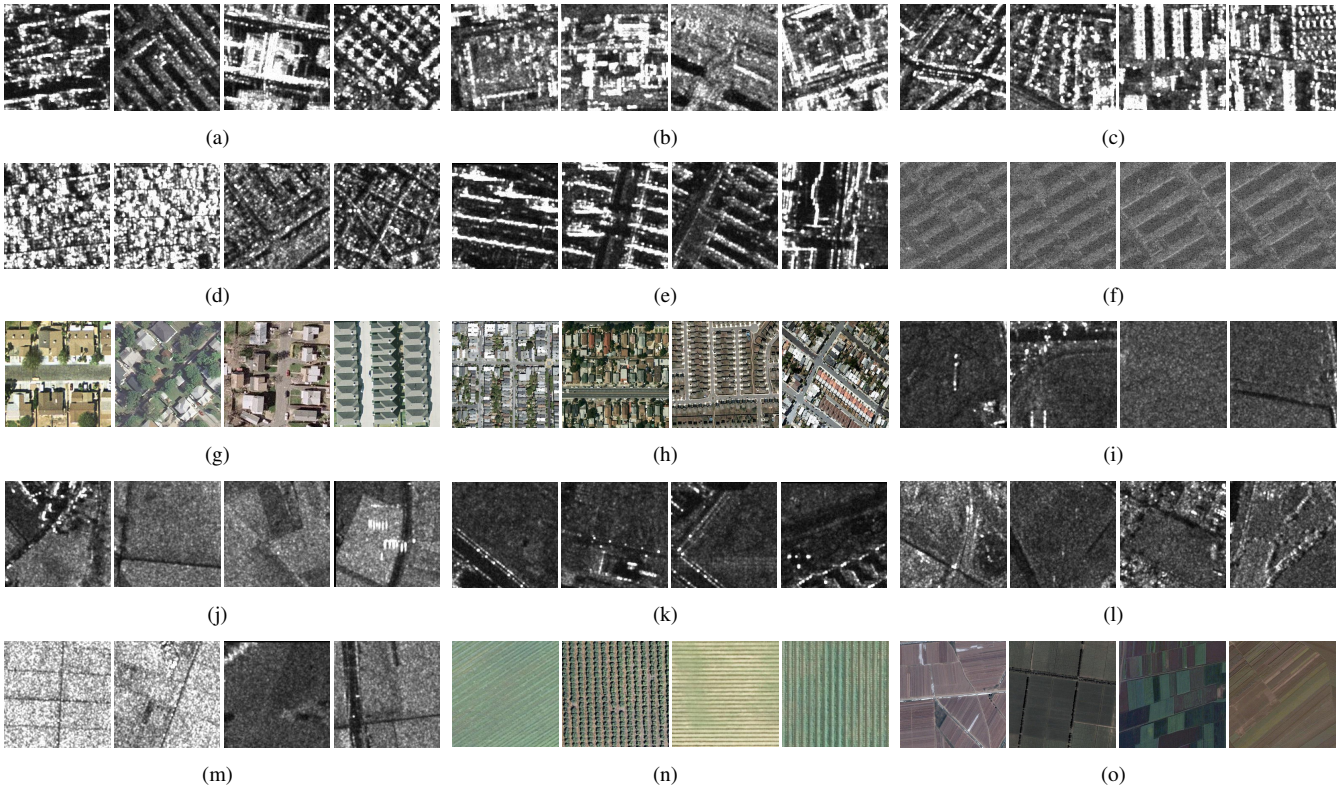


Fig. 1. Example patches corresponding to the category "Urban/Residential areas" for the datasets (a) D1, (b) D2, (c) D3, (d) D4, (e) D5, (f) D6, (g) D7, (h) D8. And corresponding to the category "Agricultural fields" for the datasets (i) D1, (j) D2, (k) D3, (l) D4, (m) D5, (n) D7, (o) D8.

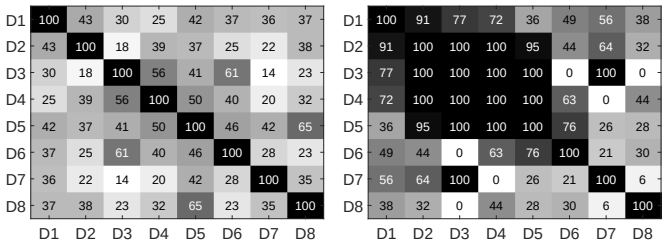
patch content, as opposed to using preselected categories. Image patches often present heterogeneous image content, and this is reflected in labels such as "High density housing area-type 3" (indicating that not all high density housing areas look alike), or "Forest with different objects." Additionally, the purpose of the dataset determines which images are selected, and this limits the dataset to that purpose.

The image-based dataset generation method stands in contrast to the *"label-based"* one used for D6-D8. In this method, different categories and their corresponding labels are first selected, and then visually homogeneous image patches whose content clearly represents the label are hand-picked. In these datasets, the number of patches are usually equally distributed between categories, and labels tend to be clear-cut, such as "Grassland" and "Pond." In this way, ambiguities which could lead to divergent understandings are reduced, therefore decreasing the effects of the sensory gap. Since the semantic gap builds on the sensory gap [6], it is also reduced. However, it was experimentally shown in [6] that these gaps are mostly reduced for the annotator and the specific purpose of that dataset (which guides the pre-selection of the labels), and therefore semantic and sensory biases are still present in the dataset. Additionally, the categories in these datasets tend to be highly homogeneous, and usually not reflecting the complexity of acquired remote sensing images.

D5 corresponds to high-resolution SAR data, and its generation method finds a middle ground between these image-based and label-based methods. D5 was created based on existing annotated SAR data; additional categories were then defined and image patches containing the appropriate content were added.

A. Experimental Procedure

The eight datasets from Table I are compared to understand their semantic content intersections. The co-occurence of categories with synonymous semantic labels was calculated between datasets. We refer to this as a semantic content intersection. For example, D1 has different labels for urban



(a) (b)

Fig. 2. (a) Semantic content intersection between datasets. (b) Percentage of exact label matches within the intersected semantic content. The numbers on the axes correspond to the different datasets as listed in Table I. The numbers inside the matrices refer to percentages.

and housing areas, such as "High density housing area type 1" and "High density urban area," while D6 has the label "Urban." These labels are synonymous, and as depicted in Fig. 2 (a), 37% of the categories within these two datasets have intersecting semantic content. Considering only the intersected semantic content, Fig. 2 (b) shows the co-occurence of exactly matching labels (category names). Referring back to the example of D1 and D6, of the 37% intersecting semantic content, only 49% are an exact match between the assigned labels.

B. Results and Discussion

Referring to D3 and D6 in Fig. 2, the effects of subjectivity in labeling image content is highlighted. D3 and D6 share a quite high semantic content intersection; however, their exact label match is null. This could be attributed to the dataset generation method: D3 followed an image-based method, whereas D6 followed a label-based method.

Datasets created with an image-based method (D1-D4) are likely to have a wider diversity of labels to describe the full image content. Therefore, it is probable that a higher degree of intersection exists between their semantic content. Considering datasets created with a label-based method (D6-D8), the opposite effect is evident. A reduced semantic content intersection between the datasets is a consequence of categories having been selected to tailor the dataset for a specific purpose.

In Fig. 2 (b), it is possible to note that overall, D1-D4 have a high percentage of exact label matches. These datasets were labeled by the same annotator, indicating that a consistent personal idea of semantics is applied across images. This consistency demonstrates the personal subjectivity in the semantic understanding of image content.

It is also possible to note that D5's semantic content, and particularly the percentage of exact label matches, are overall quite similar to datasets D1-D4. This result is not surprising, considering that D5 was created based on elements of datasets D1-D4, as mentioned in Section II. Additionally, D1-D5 were created by the same research team, reflecting not only individual subjectivity, but also the influence of existing semantic structures in the research team, since colleagues build on each others' work. These results indicate that the dataset generation method has an influence on the categories selected and the semantic labels assigned to them.

III. Dataset Discriminability

In this section, we analyze the degree of shared features between datasets to visualize how datasets can be discriminated from each other, regardless of their categories. This allows us to assess each dataset in terms of how the model trained on one of them could be generalized to the other datasets.

A. Experimental Procedure

We represent the image patches within each dataset with 3 different feature descriptors: Bag-of-Words (BoW), Weber Local Descriptors (WLD), and Gabor.

BoW provides a compact representation of images through the extraction of local image features by vectorizing a sliding window of 3×3 pixels [21]. A dictionary of visual words is generated by applying a clustering method (e.g., k-means) to a sample set of the local feature vectors (1% of all the feature vectors), where the cluster centers represent the visual words. Using this dictionary, each image patch is represented with a histogram of the visual words, by assigning its local feature vectors to the nearest cluster center. In our experiments, the dictionary is created for each dataset separately.

WLD feature descriptors are constructed by a two-dimensional histogram of: 1) *Differential Excitation*, the brightness difference ratio between a pixel x and its neighbors; and 2) *Orientation*, which is the gradient orientation of a pixel x [22]. The two-dimensional histogram is quantized to M excitations and T orientations, and then built into a one-dimensional histogram, resulting in the final feature vector. In our experiments, feature descriptors are globally computed, and we set M and T equal to 6 and 8, respectively (based on [22]), which results in a feature vector of 144 elements.

Gabor feature descriptors are acquired by filtering the image with Gabor filters [23], which are linear band-pass filters that are generated through the scaling and rotation of a mother wavelet filter whose impulse responses are 2D modulated Gaussian functions. The Gabor feature vectors are then constructed through the computation of means (μ_{sr}) and standard deviations (σ_{sr}) of the response for S scales and R rotations, $F_{Gabor} = [\mu_{11}\ \sigma_{11}\ \mu_{12}\ \sigma_{12}\ ...\ \mu_{SR}\ \sigma_{SR}]$. In our experiments, image patch features are globally extracted. The selection of $S = 3$ and $R = 6$ results in feature vectors of 36 elements.

For each dataset, we randomly sample a maximum of 100 image patches from each category. If a category includes less than 100 image patches, we take the entire category. We conduct various experiments by taking different numbers of training samples (16, 32, 64, 128) from the sample images. These samples are used to train an 8-way k-nearest neighbors (k-NN) classifier [24]. The classifier is tested on 400 random image patches, taken from the sampled image patches of each dataset. The classification is repeated 10 times with different randomly sampled image patches for training and testing.

B. Results and Discussion

We classify the different image patches, as an indicator of how distinct the datasets they belong to are. Fig. 3 (a) shows the classification accuracy of the datasets when varying

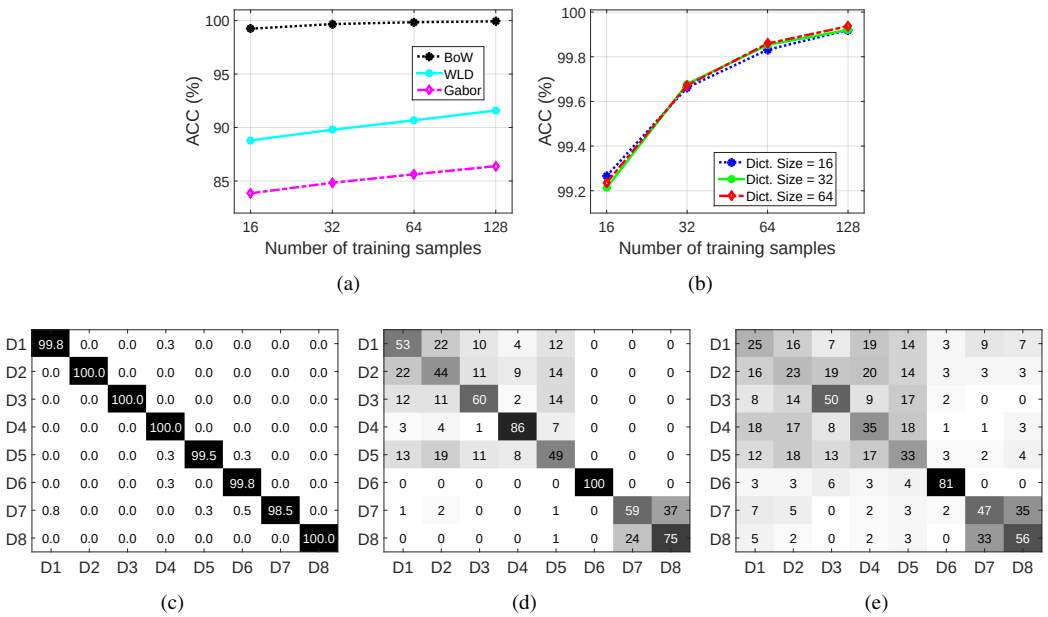


Fig. 3. (a) Classification accuracy of the datasets when varying the number of training samples. (b) Classification accuracy of the datasets using BoW features for different dictionary sizes. (c) Confusion Matrix of the datasets for BoW features with a dictionary size of 64. (d) Confusion Matrix of the datasets for WLD features. (e) Confusion Matrix of the datasets for Gabor features. For (c)-(e), the labels on the axes correspond to the different datasets as listed in Table I. The numbers inside the matrices refer to percentages.

the number of training samples for BoW, WLD, and Gabor feature descriptors. For BoW, the dictionary size was picked experimentally, by performing a classification for different dictionary sizes (16, 32, 64), as represented in Fig. 3 (b). Since the results are about the same for all dictionary sizes, we take 64 for our experiments. As the results show in Fig. 3 (a), BoW discriminates the datasets surprisingly well, even with a small number of training samples. WLD and Gabor also perform rather well (with classification accuracies higher than 80%) across the different number of training samples. From a feature perspective, the datasets are very discriminable; in other words, they have only a small feature overlap.

This is further demonstrated in the confusion matrices showing the feature overlap between the different datasets for the three feature descriptors as depicted in Fig. 3 (c)-(e). All three matrices display a pronounced diagonal, meaning that each dataset is unique to some degree. Fig. 3 (c) shows that the features obtained by BoW for each dataset are significantly different from the others. Fig. 3 (d) and (e) show that features obtained by both WLD and Gabor discriminate the datasets across sensors: the SAR datasets (D1-D6) are distinct from the optical ones (D7-D8). Within each sensor group there is also a certain degree of shared features. This pattern is more pronounced for Gabor feature descriptors.

As shown in Fig. 2 (a), D6 has a semantic content intersection with the other datasets, particularly with data from the same sensor. However, it is almost not sharing features with other datasets, as displayed in Fig. 3 (c)-(e). Within datasets D1-D4, there is a greater semantic content intersection, compared with their shared features. Although, according to Fig. 2, these different datasets have intersecting semantics, each similar label is associated with different features, which is evidence of the sensory and semantic biases in the datasets. The dataset generation methodologies cause these biases. In the case of D6, a label-based approach was used, resulting in visually homogeneous image patches within each class (Fig. 1 (f)), which are distinct from the other SAR datasets (Fig. 1 (a)-(e)). For D1-D4, an image-based approach was used. Although the datasets were created by the same annotator, similar labels refer to patches with different features. This can be seen in Fig. 1 (a)-(d), where all image patches refer to the label "Urban/Residential areas," however, the objects within the patches look very different.

Referring to Fig. 3 (d)-(e), D1 and D2 are less discriminable than the other SAR datasets for both WLD and Gabor feature descriptors. This demonstrates that their features are more generalizable over the other SAR datasets. Altogether, this shows that the datasets are distinct to some degree, and therefore the model trained on one of the datasets could not be easily generalized to the others.

IV. Cross-Dataset Generalizability

Datasets are widely used for the training and validation of content-based image retrieval methods; therefore, their built-in sensory and semantic biases can affect the resulting models. In this section, we assess the generalizability of the different remote sensing image datasets as training data, addressing the question of whether a model trained on one dataset is easily generalizable to another.

A. Experimental Procedure

In order to analyze the cross-dataset generalizability of the eight datasets for a particular object category, we trained an object detector on each dataset. The object detector is performing one versus all classification for a particular object category, using the k-NN algorithm.

BoW features										
	D1	D2	D3	D4	D5	D6	D7	D8	Drop SS	Drop DS
D1	**50.2**	17.9	0.0	28.8	27.6	0.0	6.2	0.0	70.1%	93.8%
D2	1.5	**57.1**	39.7	46.2	23.0	46.4	13.6	12.7	45.1%	76.9%
D3	0.0	31.1	**66.2**	59.2	38.2	45.7	18.3	10.0	47.4%	78.6%
D4	8.5	29.9	44.6	**85.6**	39.5	31.8	16.6	5.4	63.9%	87.1%
D5	7.1	31.1	30.1	38.7	**94.8**	24.2	18.8	11.3	72.3%	84.1%
D6	9.5	30.4	5.6	57.8	21.7	**96.8**	20.6	6.4	74.2%	86.1%
D7	0.0	3.5	0.0	0.0	0.0	0.0	**64.3**	0.0	100.0%	98.9%
D8	0.0	0.0	0.0	0.0	0.0	0.0	7.5	**55.0**	86.4%	100.0%

WLD features										
	D1	D2	D3	D4	D5	D6	D7	D8	Drop SS	Drop DS
D1	**40.1**	14.7	8.8	2.7	17.4	42.7	1.3	0.0	56.9%	98.4%
D2	14.1	**56.4**	46.6	62.7	25.5	36.5	1.2	0.0	34.2%	98.9%
D3	9.8	38.5	**66.7**	73.4	48.6	32.2	0.0	0.0	39.2%	100.0%
D4	7.9	36.7	17.6	**81.9**	25.8	31.8	0.0	0.0	70.7%	100.0%
D5	13.2	40.9	65.8	42.4	**93.1**	0.0	0.0	0.0	65.1%	100.0%
D6	0.0	0.0	0.0	0.0	0.0	**96.8**	0.0	0.0	100.0%	100.0%
D7	0.0	0.0	0.0	0.0	0.0	0.0	**67.7**	28.1	58.4%	100.0%
D8	0.0	0.0	0.0	0.0	0.0	0.0	29.3	**64.8**	54.8%	100.0%

Gabor features										
	D1	D2	D3	D4	D5	D6	D7	D8	Drop SS	Drop DS
D1	**40.5**	16.1	18.1	13.7	13.9	28.6	0.0	0.0	56.3%	100.0%
D2	12.0	**55.5**	45.7	46.7	35.0	41.6	0.0	9.1	34.7%	91.8%
D3	16.7	39.8	**64.4**	40.4	46.4	27.0	0.0	0.0	47.1%	100.0%
D4	12.9	40.0	43.8	**77.1**	42.3	53.6	0.0	0.0	50.0%	100.0%
D5	10.4	36.9	51.3	45.1	**79.8**	17.9	0.0	0.0	59.5%	100.0%
D6	9.6	16.6	38.1	46.7	36.9	**96.1**	0.0	0.0	69.2%	100.0%
D7	0.5	0.1	0.0	0.0	3.3	0.0	**60.1**	0.0	100.0%	98.9%
D8	12.5	1.9	0.0	14.5	25.8	4.1	30.2	**42.2**	28.3%	76.8%

TABLE II
CROSS-DATASET DETECTION OF THE CATEGORY "URBAN/RESIDENTIAL AREAS." THE VERTICAL AXIS REFERS TO THE TRAINING DATASETS, AND THE HORIZONTAL AXIS REFERS TO THE TRAINING DATASETS.

BoW features										
	D1	D2	D3	D4	D5	D6	D7	D8	Drop SS	Drop DS
D1	**76.3**	0.0	0.0	0.0	0.0	–	11.5	11.3	100.0%	85.0%
D2	0.0	**68.6**	0.0	0.0	0.0	–	6.9	0.0	100.0%	94.9%
D3	0.0	0.0	**73.3**	0.0	0.0	–	2.0	0.0	100.0%	98.6%
D4	0.0	0.0	0.0	**74.2**	0.0	–	9.1	11.9	100.0%	85.8%
D5	0.0	0.0	0.0	0.0	**94.3**	–	20.7	3.1	100.0%	87.4%
D6	–	–	–	–	–	–	–	–	–	–
D7	0.0	21.1	0.0	29.5	0.0	–	**66.1**	11.0	83.3%	84.7%
D8	0.0	0.0	0.0	0.0	0.0	–	0.0	**65.6**	100.0%	100.0%

WLD features										
	D1	D2	D3	D4	D5	D6	D7	D8	Drop SS	Drop DS
D1	**65.7**	54.8	13.3	0.0	44.6	–	4.4	11.9	57.1%	87.6%
D2	41.8	**70.5**	12.5	0.0	47.6	–	5.4	11.1	63.8%	88.2%
D3	27.2	37.6	**73.4**	11.0	52.8	–	4.6	10.9	56.2%	89.4%
D4	10.2	24.1	29.9	**70.9**	49.2	–	5.2	10.1	60.0%	89.2%
D5	21.5	48.4	36.8	45.2	**85.5**	–	4.8	10.4	55.6%	91.1%
D6	–	–	–	–	–	–	–	–	–	–
D7	0.0	0.0	0.0	0.0	0.0	–	**86.3**	6.1	92.9%	100.0%
D8	0.0	0.0	0.0	0.0	0.0	–	6.4	**54.1**	82.6%	100.0%

Gabor features										
	D1	D2	D3	D4	D5	D6	D7	D8	Drop SS	Drop DS
D1	**57.8**	17.5	13.3	12.8	16.9	–	12.6	0.0	73.8%	89.1%
D2	31.3	**60.6**	19.3	14.5	56.9	–	0.5	0.0	49.6%	99.6%
D3	24.2	11.2	**61.2**	37.1	18.1	–	9.3	8.5	62.9%	85.4%
D4	17.1	27.2	36.3	**71.3**	54.1	–	1.5	0.0	52.7%	98.9%
D5	22.3	45.1	22.8	34.8	**77.2**	–	1.1	0.0	59.5%	99.2%
D6	–	–	–	–	–	–	–	–	–	–
D7	0.8	0.0	0.0	0.0	0.0	–	**95.6**	19.6	79.5%	99.8%
D8	3.8	0.0	0.0	0.0	0.0	–	42.7	**72.7**	41.3%	98.7%

TABLE III
CROSS-DATASET DETECTION OF THE CATEGORY "AGRICULTURAL FIELDS." THE VERTICAL AXIS REFERS TO THE TRAINING DATASETS, AND THE HORIZONTAL AXIS REFERS TO THE TRAINING DATASETS.

We then use the resulting model to detect that object category in all datasets. For each category, 30% of the image patches are used for training, and the rest are used for testing. The image patches are represented by BoW, WLD and Gabor feature descriptors.

B. Results and Discussion

Tables II and III refer to the cross-dataset generalizability assessment for two commonly used categories: "Urban/Residential areas" and "Agricultural fields." These categories were selected because their image patch content is visually different. Image patches corresponding to the "Urban/Residential areas" category are highly structured, whereas the "Agricultural fields" category is more visually homogeneous. Tables II and III are made up of three sub-tables, one for each feature descriptor (BoW, WLD, Gabor). In each sub-table, the vertical axis refers to the training datasets, and the horizontal axis refers to the dataset the model was tested on. Each sub-table has two additional columns: *Drop SS* and *Drop DS*. Both these values show the percentage decrease (Drop) when training on one dataset and testing on the others, compared with training and testing on the same dataset. For computing Drop SS, we average over the detection performance on the datasets from the same type of sensor (SS) as the training dataset (SAR or Optical). For Drop DS, we average over the detection performance on the datasets with different sensor types (DS; SAR vs. Optical) as the training dataset.

The detection performance is measured using the F-measure:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{1}$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$.
TP (True Positive) and FP (False Positive) correspond to the number of correctly and incorrectly labeled objects, respectively. FN (False Negative) corresponds to the unlabeled objects.

As we can see in Tables II and III, when we train on one dataset and test on others, there is a significant decrease in detection performance. This decrease is largest when we train on data from one sensor type and test it on data from another sensor type (Drop DS). Referring to the tables for BoW for both categories, with either homogeneous patch content such as "Agricultural fields," or with structured patch content such as "Urban/Residential areas," the performance decrease is larger than for WLD and Gabor. This is consistent with the discriminability shown in Fig. 3 (c)-(e). In the case of SAR dataset categories with fully developed speckle (e.g., "Agricultural fields"), the cross-dataset detection performance decreases by 100%. This can be explained by the fact that BoW features are highly localized (using local windows), and therefore the classes with fully developed speckle [17] result in a more unique signature. Consequently, for categories with these characteristics, the model trained on BoW features of one SAR dataset can not be further applied to other SAR datasets.

Considering only the optical datasets, the models trained

on WLD and Gabor are more generalizable for categories with structured patch content; whereas for categories with homogeneous patch content, models trained on Gabor are more generalizable.

Based on the findings from the current and previous section, due to the intrinsic biases, these datasets are distinct from each other. Therefore, training or validation of content-based image retrieval methods on one dataset does not necessarily mean that they will generally perform well.

V. Creating Less Biased Datasets

Consistent with previous research [1], the results above indicate that each dataset comes with its own bias. This raises the question: *How to create datasets which are less biased and more representative of the real world?*

Torralba and Efros [1] recommend image sampling from multiple sources in a random manner, as opposed to hand-picking images. The latter approach results in datasets which perform worse in terms of cross-dataset generalizability. This is consistent with our findings regarding datasets created with a label-based approach (D6-D8). These datasets have less semantic content intersection as shown in Fig. 2, as well as fewer shared features, depicted in Fig. 3 (c)-(e).

Furthermore, the authors [1] also recommend crowdsourcing, an approach that is used extensively in various research fields since a number of years [25], [26], [27], including computer vision, where it is considered to be a promising procedure for creating datasets [8].

Crowdsourcing's strength comes from consensus, which is directly related to the number of contributions. However, large scale public crowdsourcing is not always possible. For example, with certain high-resolution remote sensing images, usually annotations by experts are relied on as a reference. This method of dataset creation is limited by the relatively small number of experts available, and the large amount of effort necessary for creating large scale datasets. Furthermore, expert annotators can bias a reference dataset through errors [13], and subjective understandings [6]. Consequently, not many large scale remote sensing image datasets exist, and the existing ones suffer from intrinsic biases, as discussed in the previous sections. Therefore, we study the effects of different image annotation task structures (collaborative vs. individual annotation) on user annotation performance, behavior, and image understanding. This will lead us to design annotation task structures which produce more reliable image annotations, which are more likely to be accepted by other users. This in turn has implications for the creation of less biased image datasets.

According to [7], the specific characteristics of remote sensing images makes their annotation a more challenging task for users, which in turn magnifies the effects of the annotation task structure on the results. First of all, the sensors used to capture the images vary greatly (e.g., SAR, optical), therefore their features do as well. Particular objects can be better detected with certain types of sensors, while resolution and scale affect the semantic level at which the user can identify objects in an image. Additionally, the human eye is not accustomed to the perspective view from above contained in satellite images, making object identification harder. This view from above also creates an image where there is no clear foreground object distinguished from the background. Depending on the incidence angle and pass direction, objects can appear differently [28]. Due to seasonal changes, even the time of year has an impact on the image interpretation of remote sensing images. Winter scenes present different information compared with summer scenes.

Additionally, there is the issue of semantics, remote sensing scenes can be labeled at multiple semantic levels (e.g., "House," "Neighborhood," "Urban area"). The amount of context available to the user is determined by the size of the image presented to the user (e.g., an entire image or just a small part of it). Previous research with remote sensing images [2] found that when presenting users a full remote sensing scene to annotate, they tended to assign only higher level semantic labels, such as "Urban" or "Industrial area," as opposed to lower level semantics, such as "House" or "Factory." On the other hand, when users were given small parts of the scene, the labels assigned corresponded to lower level semantics.

In our experiments, we vary the annotation task structure by having non-expert users work either individually or collaboratively (pairwise) on an image annotation task. See *et al.* [14] compared the annotation quality of crowdsourced data between experts and non-experts on a discrimination and an identification task using the Geo-Wiki platform. The discrimination task had annotators assess the degree of human impact, and the identification task had them identify the type of land cover in different locations across the world. The authors found that in discrimination tasks, there was no significant difference in the performance of experts compared with non-experts. For the identification tasks, the experts initially performed better than the non-experts; however, over time non-experts' performance equaled and in some cases even outperformed the experts. These results indicate that non-experts can be used for annotation tasks without a decrease in data quality. Although the number of non-experts available to annotate remote sensing datasets is greater than the number of experts, it is also somewhat limited due to the constraint of not being able to use large-scale online crowdsourcing for certain types of remote sensing data (e.g., non-public high resolution multispectral and SAR images). Therefore, the annotation task structure must be optimized to ensure that non-experts can efficiently produce accurate and thorough annotations, with a limited number of annotators.

A. Experimental Procedure

The process chain followed for these experiments is summarized in Fig. 4. First, a multispectral *Scene* of the north of Munich (Germany) was selected. The scene was acquired on July 12th, 2010 (10:30 am UT) by the WorldView-2 satellite, and trimmed to 2000×1800 pixels. The image has a resolution of 1.84 m, and 3 bands (RGB) were selected and displayed. This scene was cut into patches, each comprising 200×200 pixels, with 50% overlap, producing 323 image patches.

In the *Initial user annotation*, three different users were each given an average of 108 image patches to annotate, meaning they had to identify the objects in the image patches, and assign them labels. The users performed a free text annotation (without a dictionary or reference) [29]. After removing duplicates and synonyms in the *Label collection & refinement* step, a dictionary of 18 *Content labels* was created (please refer to Table IV). This dictionary was then used in a manual annotation of the scene, using Google Earth[3] as a ground truth to produce a *Reference annotation* (REF). The REF is used for further analysis of the user annotations.

Thirty users were then recruited to participate in a set of user experiments, which we refer to as UX_B. To explore the effects of collaboration on the image annotation task, users were placed in either a collaborative or in an individual condition. Each individual, or each pair of users were given 53 image patches to label using the content label dictionary (please refer to Table IV). Twelve users were placed in the collaborative condition (UX_B_C), performing the annotation task with a partner. These six pairs each annotated the same 53 image patches. The remaining 18 users were assigned to the individual condition. Of these 18 users, six of them annotated the same image patches as UX_B_C, we refer to this group as UX_B_{MI}. This was done so that the image patches were held constant, and we could compare the labels assigned by the individuals and the collaborators. The other 12 users in the individual condition are referred to as the UX_B_{IR} group. Each user in this group annotated different sets of 53 randomly selected image patches. This group allows us to compare the performance of individuals when we vary the image patches given. Information on the different experimental conditions can be found in Table V.

After users were assigned to the different conditions, they were given a short demo of the annotation tool. Please refer to Fig. 5 for a screenshot of the tool. The tool has zoom and panning functions, as well as 8 drop-down menus displaying the labels from Table IV. Users were asked to look at each image patch, and assign labels from the drop-down menus to describe the objects they could see, in whatever order they saw them. Users were additionally told to take their own interpretation of the label meanings. Users were given as much time as they required to finish the annotation of the given image patches. The quantitative data collected were the labels assigned to the image patches.

In addition to this, qualitative data was collected in the form of semi-structured individual interviews with the users, which took place after the annotation task was completed. The 11 interview questions sought to explore the users' experience with the task, including why the user would, or would not, want to continue with the annotation task, what aspects of the annotation task were easy or difficult, why this was the case, and asked users to rate their confidence in their annotations on a scale, among others. These questions were used to launch into a dialog with the user, providing insights into his/her subjective perceptions of the annotation task, such as which objects were more difficult to label, why this was the case, problems they ran into while annotating, and how they approached the annotation task.

[3]https://www.google.com/earth/

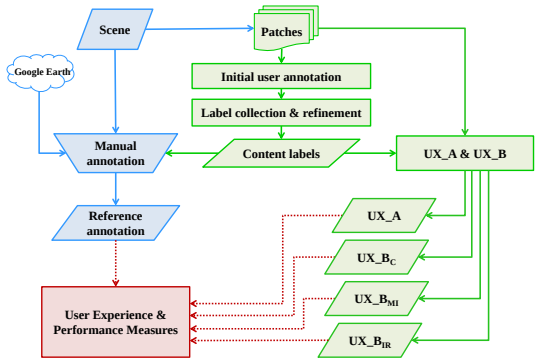


Fig. 4. Process chain of our user experiments

1	Agricultural field	7	Greenhouse	13	Railway
2	Building	8	Highway	14	Road
3	Crop	9	House	15	Soccer field
4	Factory	10	Isolated trees	16	Solar panels
5	Forest	11	Lake	17	Street
6	Grass	12	Parking lot	18	Tennis court

TABLE IV
CONTENT LABELS

B. Results and Discussion

In the following subsections we analyze the quantitative and qualitative data collected, to further understand how the different scenarios and conditions affected different aspects of user annotation performance.

1) User Overall Performance: We calculated average performance measures across different experimental conditions, which is presented in Table VI. Overall, UX_B_C have the highest average Precision, Recall, and F-measure. The overall variance measures show that UX_B_C's results are less distributed. This is also demonstrated in Fig. 6 (b), where the points representing the performance of UX_B_C are closely grouped together, as opposed to the points representing the performance of UX_B_{MI} and UX_B_{IR}, which are distributed over a larger range. This figure additionally shows that the individuals' variance is larger for Recall, therefore the main difference between the different individuals' performance is the number of objects they could identify, and not the precision of the assigned labels. In the case of UX_B_C, not only is the user performance closer together, but so are the labels assigned, which is reflected in the user agreement. According to the results, the user agreement among UX_B_C is 79.6%, while among UX_B_{MI} is 65.2% (only these 2 groups were considered because they labeled the same image patches). As indicated in Table VI, UX_B_C's higher performance does not require additional time, as compared to UX_B_{MI} and UX_B_{IR}.

2) User Confidence Ratings: Fig. 6 (b) shows Precision versus User Confidence ratings. As we can see, the users in the individual condition (UX_B_{MI}, UX_B_{IR}) give a conservative estimate of their precision. The users in the collaborative condition (UX_B_C) are overall better at estimating their precision, and show less variance in their ratings. The interviews

User Groups	Sample Size (# users)	Image Patches Labeled	Quantitative Data Collected	Qualitative Data Collected	Condition
UX_B$_C$	12 (6 pairs)	53	Object labels	Interview	Collaborative: working in partners
UX_B$_{MI}$	6	53	Object labels	Interview	Matched individuals: individuals labeled same image patches as UX_B$_C$
UX_B$_{IR}$	12	53	Object labels	Interview	Rest of the individuals: labeled different image patches

TABLE V
OVERVIEW OF THE DIFFERENT CONDITIONS IN THE USER EXPERIMENTS.

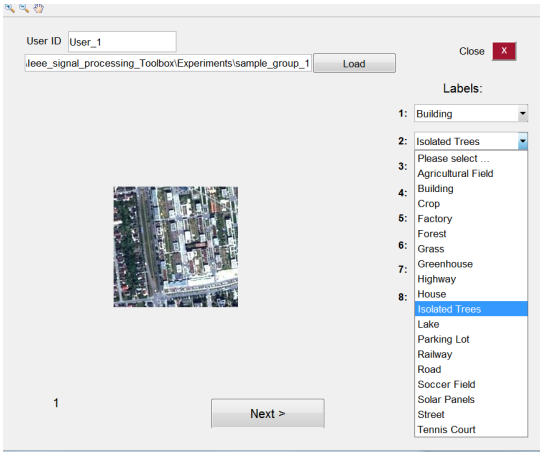


Fig. 5. Screenshot of the annotation tool used for the user experiments

	UX_B$_C$	UX_B$_{MI}$	UX_B$_{IR}$	Users in individual condition (UX_B$_{MI}$ + UX_B$_{IR}$)
Avg. Precision (%)	**88.05**	82.58	82.55	82.67
Avg. Recall (%)	**66.28**	56.48	53.27	54.34
Avg. F-measure (%)	**75.42**	65.51	63.26	64.01
Variance Precision	**21.30**	42.90	3.98	16.98
Variance Recall	**17.76**	244.58	233.91	239.75
Variance F-measure	**3.83**	189.11	199.12	196.91
Avg. Time/Patch (sec.)	**44**	53	42	47.5

TABLE VI
AVERAGE PERFORMANCE MEASURES

reveal that when giving confidence ratings, users consider the correctness of their assigned labels (they are not considering the objects they may have missed).

3) Class-wise User Performance Analyses: Fig. 7 presents Precision, Recall, and F-measure for different experimental conditions, across the 18 labels shown in Table IV. The blue, red, green, and yellow bars represent UX_B$_C$, UX_B$_{MI}$, UX_B$_{IR}$, and the average over all of UX_B conditions, respectively. Fig. 7 (c) shows that the biggest difference in performance between the UX_B$_C$ and the individuals (UX_B$_{MI}$ and UX_B$_{IR}$) is for the classes that were considered as hard to identify based on the user interviews. Taking the example of "Railway" and "Parking lot," user interviews describe these classes as difficult to identify due to image product properties, namely resolution and scale. This is reflected in Fig. 7 (b), where the low Recall shows that users had trouble recognizing these objects. Other ones, such as "Street," were also hard to identify due to semantic confusion with classes such as "Road." Based on a calculation of the confusion between classes, UX_B$_C$ tend to confuse "Street" with "Road" in 14% of the cases; whereas for individuals this is true for 28% of the cases.

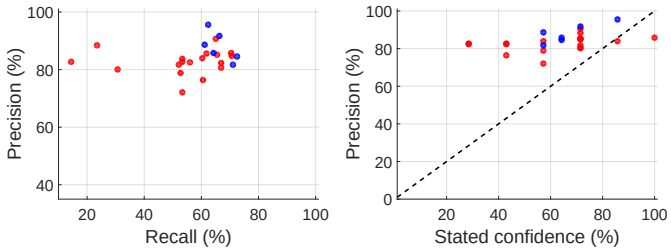


Fig. 6. (a) Precision vs. Recall. (b) Precision vs. Users' stated confidence. The red points in the graphs refer to UX_B$_{MI}$ and UX_B$_{IR}$, whereas the blue dots refer to UX_B$_C$.

4) Object Visibility for Users: Users were told to assign up to eight labels to each image patch following the order in which they identified them, allowing us to measure the visibility as the average number of times the label was assigned to each of the eight possible positions. Fig. 8 shows the visibility of each label, comparing UX_B$_C$ and UX_B$_{MI}$. Here we can see that "Agricultural field" and "Lake" are the most visible object classes for all users, with this label being assigned in first place more than 50% of the time. Since these object categories were mostly co-occurring, it is possible to see how the labeling order between them varied for UX_B$_C$ compared to UX_B$_{MI}$. For UX_B$_{MI}$ there is a larger probability that "Agricultural field" was labeled first, whereas for UX_B$_C$ there is a larger probability that "Lake" was labeled first. "Crop," "Forest," and "Highway" also co-occur with the two previously mentioned categories, and here we can also see differences between UX_B$_C$ and UX_B$_{MI}$. For UX_B$_C$, the categories "Lake," "Forest," and "Highway" are usually labeled prior to the categories "Agricultural field" and "Crop." This pattern may reflect that users in the collaborative (UX_B$_C$) condition first labeled the categories both users were sure of and agreed upon, leaving the ones they had to discuss to reach an agreement till the end.

User interviews indicate that of these categories, confusion exists between the two categories "Agricultural field" and "Crop," as well as between "Road" and "Street." In an average of 4.5% of the cases, all users misassign the label "Crop" to agricultural fields; whereas confusion in the other direction occurs in 55% of the cases. Among the other two labels, the label "Road" is miassigned to streets in an average of 21%, whereas confusion in the other direction occurs in 25% of the cases. This can justify the difference between the visibility distribution of these two pairs of labels. "Agricultural field" and "Crop" display a visibility distribution with a strong preference for the first positions. "Road" and "Street," on the other hand, present a more even distribution across positions. This

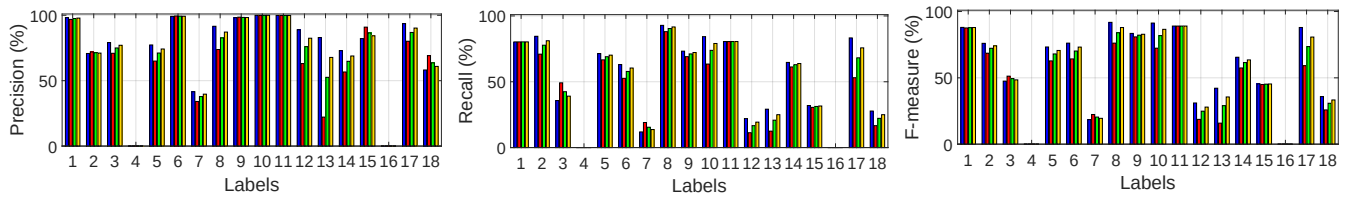


Fig. 7. Class-wise performance for the 18 labels shown in Table IV. The blue, red, green and yellow bars represent UX_B_C, UX_B_{MI}, UX_B_{IR}, and the average over all UX_B conditions, respectively.

can be further explained by the differences in confusion. In the majority of the cases "Crop" is confused with "Agricultural field", whereas confusion is bi-directional between "Road" and "Street". Considering that the co-occurences of both pairs of object categories are similar (39% for "Agricultural field" and "Crop," and 33% for "Road" and "Street"), their visibility distribution is not due to co-occurence, but rather to sensory and semantic confusions between object categories.

5) User Performance Measures Throughout the Task: User performance measures are plotted against Patch ID which reflects the order that users saw the 53 patches, giving us some insight into how users acquire experience and skills, while progressing with the annotation task. In Fig. 9 (a), F-measure is higher for UX_B_C, showing that their overall performance is better. Comparing Fig. 9 (b) and (c), it is visible that Precision is closer together than Recall for UX_B_C and UX_B_{MI}, indicating that the differences in overall performance are mostly attributed to the number of unidentified objects. The lower average Recall of UX_B_{MI} shows that individuals are more likely to miss objects, either because they did not detect the object, or they were insufficiently confident of their object identification and therefore decided to leave it unlabeled instead.

These figures also show that after approximately 10 image patches, the user performance appears to stabilize in all conditions. This is confirmed in the interviews, where users express doubts regarding the accuracy of their first 10 image patch annotations. Users utilize these first patches to become familiar with the appearances of different objects, and create their own working definitions of the labels. Fig. 9 (d) plots elapsed annotation time over Patch ID, showing that all users are improving their performance as they get exposed to more patches. Additionally, users are also becoming quicker at doing the annotation, with UX_B_C becoming slightly quicker than UX_B_{MI} by the end of the annotation, and showing a slightly increasing trend.

To better understand how experience with the task affects user performance, eight classes were chosen to represent a variety of features: highly structured objects ("Building," "House"), homogeneous objects ("Agricultural Field," "Crop," and "Grass"), objects with a small coverage which are hard to identify ("Parking lot,"), and objects with distinguishing features ("Road," "Street"). For each selected label, we calculated the Recall, and the False Discovery Rate ($FDR = \frac{FP}{FP+TP}$) as an error measure, where FP (False Positive) represents the incorrect assignment of the label to another object, and TP (True Positive) represents the correct assignment of the label. These measures are plotted over Patch ID in Fig. 10. Since user performance could additionally be affected by the coverage, we measured user's Recall over the class coverage for the same eight classes. The results of these experiments are described below:

- The correct identification of the object category "Agricultural field" is highly dependent on label coverage of up to 20%, as we can see in Fig. 10 (a). Both UX_B_C and UX_B_{MI} have a very similar Recall which is overall high. This measure of performance stabilizes after the user has seen approximately ten images. In Fig. 10 (a), we can see that this label is seldomly incorrectly assigned to other categories, therefore it is very identifiable. This is also consistent with the visibility of this category, shown in Fig. 8 (a).
- The object category "Building" requires at least 5% label coverage before the performance stabilizes, as shown in Fig. 10 (b). For UX_B_{MI} the Recall at this point is lower than that of UX_B_C, meaning that the users did not correctly identify this label in more cases. Recall stabilizes after the users have seen approximately ten image patches, which is consistent with the interviews. Additionally, users expressed that they had some semantic confusion with the label "House." Based on our analysis of the annotations, in 19% of the cases, both groups of users assigned the label "House" to buildings.
- The correct identification of the object category "Crop" is not very dependent on coverage, as shown in Fig. 10 (c). Both user groups' Recall is stable across Patch ID, even though the coverage of this object class changes throughout, as depicted in Fig. 9 (e). Overall Recall for both user groups' is low, due to the sensory and semantic confusion with the object class "Agricultural field." However, it is higher for UX_B_{MI}, which can be explained by the lower misassignment of the label "Agricultural field" to crops (47% for UX_B_{MI} vs. 62% for UX_B_C). Although UX_B_{MI}'s Recall is higher than UX_B_C's, UX_B_{MI}'s error is higher, indicating that this group of users misassigned the label "Crop" to other categories more frequently.
- Although the object category "Grass" is homogeneous, unlike the categories "Agricultural field" and "Crop" it is not as visible (as shown in Fig. 8 (f)). Its identification is also not dependent on coverage (as shown in Fig. 10 (d), probably because its coverage is usually distributed throughout the entire patch in small blocks, as opposed to large continuous areas. This object category is often seen as a background object, which is also reflected

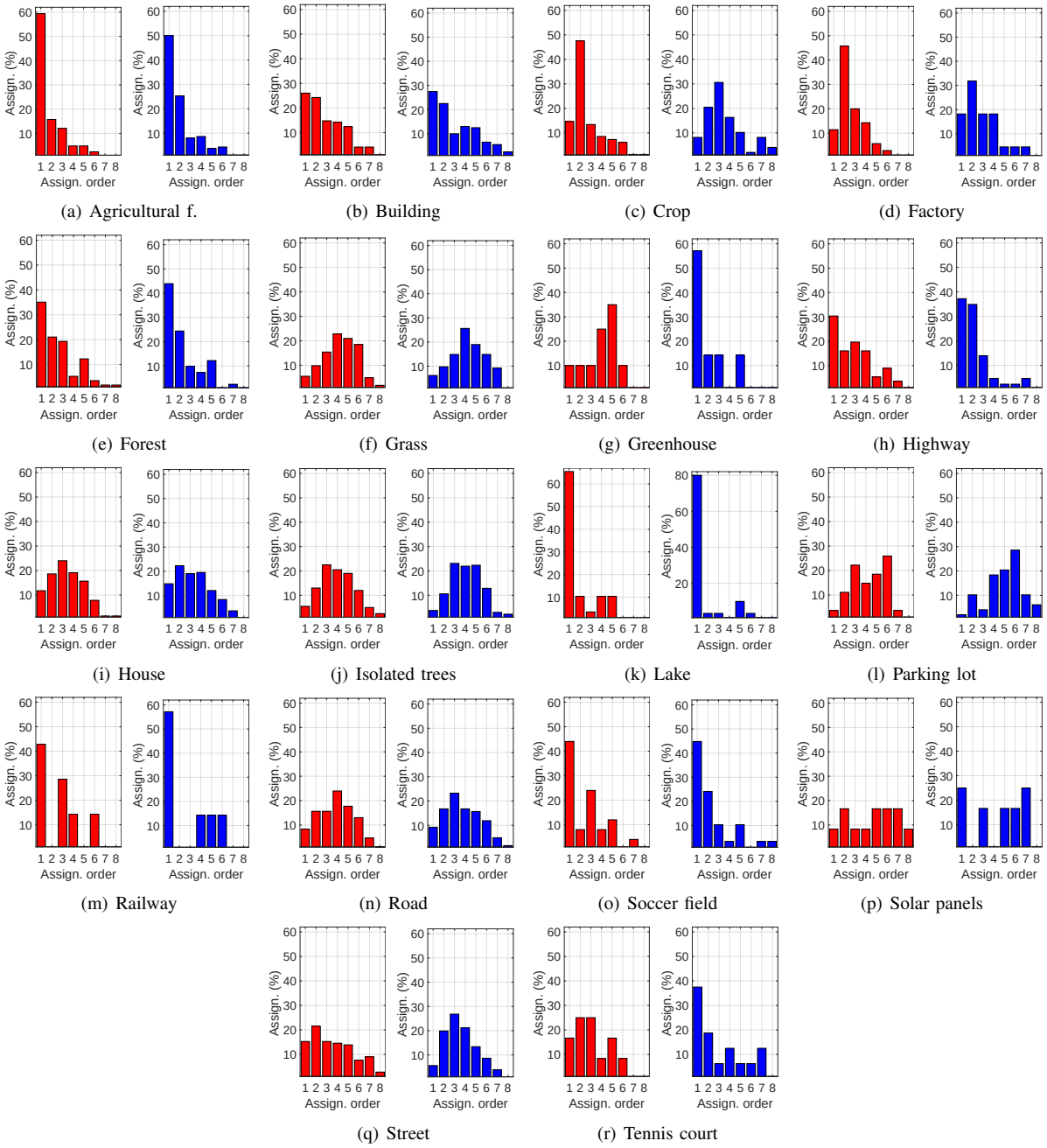


Fig. 8. The visibility of each object categories, reflected in the order in which users identified and assigned the labels. The red bars refer to UX_B_{MI}, while the blue bars refer to UX_B_C.

in its error rate. It is either correctly identified, or missed, but not incorrectly assigned as a label to other object categories.

- The correct identification of object category "House" is dependent on coverage up to about 15%, as shown in Fig. 10 (e). Fig. 9 (e) shows that between image patches 10 and 30, this object class' coverage is below 20%. At this point there is a slight decrease in Recall. After image patch 30 there is an increase in coverage, therefore after this point there is also a stabilization in performance. This label is seldomly incorrectly assigned to other categories, however in 21% of the cases, the label "Building" was assigned to houses.
- The coverage of the object category "Parking lot" is small, as it is always below 2% on average, as shown in Fig. 9 (f). Therefore, as Fig. 10 (f) shows, users' Recall is not dependent on it. User interviews indicated that it was perceived as one of the harder category to identify, which is also evident in Fig. 8 (l), where users tend to assign this label toward the end. Additionally, UX_B_{MI} wrongly assigned this label to other objects more than UX_B_C.
- Although coverage of the object category "Road" is low (as shown in Fig. 9 (e)) , Recall for all users is relatively high, indicating the correct identification of this object category is not dependent on coverage, which is also evident in Fig. 10 (g). This is probably due to its

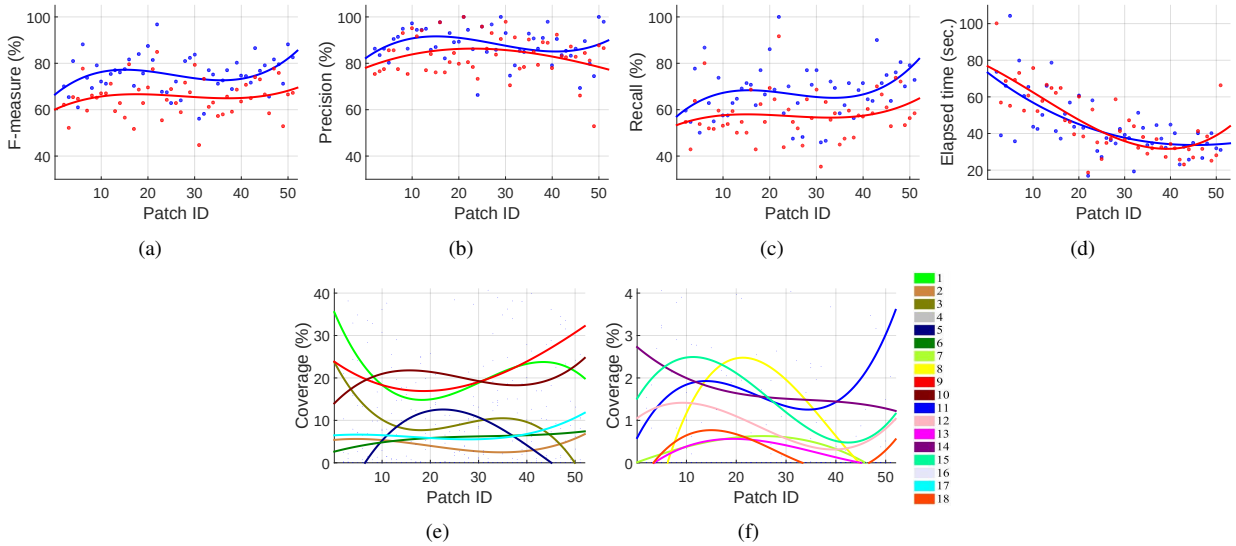


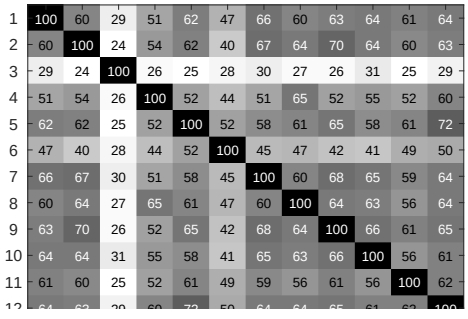
Fig. 9. Performance plotted against Patch ID. (a) F-measure, (b) Precision, (c) Recall, (d) Elapsed time. The red points refer to UX_B_{MI}, whereas the blue points refer to UX_B_C. The lines are a polynomial curve fitted to the point, and indicate the trend. (e) and (f) display the labels with average coverages over and under 4%, respectively. The area covered by each label is plotted against Patch ID for all 18 classes. The numbers in the legend refer to the labels in Table IV.

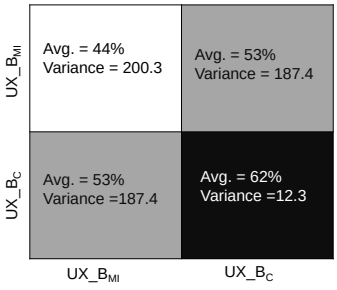
salient features which become apparent even at low label coverages. Additionally, roads tend to cross homogeneous surroundings, making them easier to identify.

- Although roads and streets share some features, streets are harder to identify because they are usually surrounded by many objects. Therefore, the correct identification of the object category "Street" is more dependent on label coverage, as shown in Fig. 10 (h). Based on the interviews, "Street" and "Road" created some semantic confusion for all users; however, according to Fig. 10 (h), UX_B_C had on average a 20% higher Recall than UX_B_{MI}. This could be due to UX_B_C's lower degree of confusion between these two labels. While UX_B_C misassigned the label "Road" to streets in 14% of the cases, for UX_B_{MI} it was in 28% of the cases.

The results above demonstrate that in general, homogeneous object categories require higher coverage compared to structured categories in order to be consistently recognized by the users. The average performance measures from Table VI indicate that collaborators (users working in pairs) perform better than individuals, and with a lower variance among their measures, which makes their performance more predictable. This also means that any given label assigned by the collaborators is more likely to be generally accepted by other users, compared to any given label from the individuals.

Furthermore, when we analyze performance by object category, we can see that this increased performance is due to the collaborators increased ability to identify the more difficult object categories. For more visible ones, like "Agricultural field," collaborators and individuals have overlapping performance. However, for the harder classes such as "Parking lot" or "Street," collaborators performance is much better than the individuals, implying that they are able to overcome some of the difficulties presented by semantic confusion (e.g., between "Road" and "Street"), and by certain product properties (e.g., scale). Additionally, in user interviews collaborators recognized the value of working with a partner and its positive impact on their results of the annotation task. Users in the individual condition were also asked to imagine what the task would be like if they had done it with a partner, with most of these users predicting that their results would be better, and additionally the task would be more enjoyable.





(b)

Fig. 11.

VI. Conclusion

Correctly annotated image datasets are important for developing image mining methods. However, there is still some doubt on the generalizability of the available datasets as training data, raising the question of whether a model trained on one dataset is easily generalizable to another dataset. In this article we experimentally demonstrate the existing dataset biases on eight different remote sensing datasets. We first assess the degree of shared features between datasets, finding that there is little feature overlap between datasets. We then assess the degree of generalizability of an image dataset as training data for a model. To this end, a model was trained and tested on one dataset, and then applied to another one, with results indicating a large performance decreases, therefore the generated models were not generalizable.

Crowdsourcing has been suggested as a methodology to overcome these issues and create less biased datasets. However, in some cases, such as remote sensing images, large scale online crowdsourcing is not always an available option in terms of structuring an image annotation task, due to the restrictions on making certain images publicly available. Therefore, we explored a collaborative image annotation task structure using a limited number of non-experts, and assessed how it influenced user behavior, performance and image understanding relative to the annotation task.

Our results indicate that collaborators outperform individuals on correctly identifying objects in images, and their results have a lower variance, making their performance more predictable. Therefore, any given label assigned by the collaborators is more likely to be generally accepted by other users, compared to any given label from the individuals. Additionally, their higher performance is mostly due to an increased ability to identify more difficult object categories. Through user interviews, we identified that collaborators were aware of the value of working with a partner. Moreover, most of the users in the individual condition reported that they would find the task more enjoyable if they could do it with a partner, and that the results would be better.

Therefore, due to the positive impact of a collaborative task structure on user performance, a collaborative task structure could be considered for designing efficient image annotation methodologies using small groups of non-experts, which produce more reliable image annotations.

Further studies are needed to assess the effects of collaborative image annotation to create annotated image datasets, and to test the generalizability of these datasets.

Acknowledgements

This work was supported in part by Munich Aerospace. The image data used in this study were provided by the TerraSAR-X Science Service System and by European Space Imaging. We would additionally like to thank the various volunteers who participated in our study, as well as our colleague G. Schwarz for providing helpful hints.

References

[1] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521–1528.

[2] R. Bahmanyar and A. Murillo Montes de Oca, "Evaluating the sensory gap for earth observation images using human perception and an LDA-based computational model," in *Proc. IEEE International Conference on Image Processing (ICIP)*, September 2015, pp. 566–570.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, December 2000.

[4] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 1, pp. 193–204, March 2011.

[5] R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 4377–4381.

[6] R. Bahmanyar, A. Murillo Montes de Oca, and M. Datcu, "The semantic gap: An exploration of user and computer perspectives in earth observation images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2046–2050, October 2015.

[7] A. Murillo Montes de Oca, N. Nistor, and M. Datcu, "Creating a Reference Data Set for Satellite Image Content Based Retrieval," in *Proc. Conference on Big Data from Space (BiDS)*, Frascati, 2014, pp. 71–75.

[8] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2008, pp. 1–8.

[9] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, May 2008.

[10] L. von Ahn and L. Dabbish, 2004, pp. 319–326.

[11] ——, *Communications of the ACM*, vol. 51, no. 8, p. 57, August 2008.

[12] R. Werninghaus and S. Buckreuss, "The TerraSAR-X mission and system design," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 2, pp. 606–614, Feb 2010.

[13] G. M. Foody, L. See, S. Fritz, M. Van der Velde, C. Perger, C. Schill, and D. S. Boyd, "Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project," *Transactions in GIS*, vol. 17, no. 6, pp. 847–860, 2013.

[14] L. See, A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner, "Comparing the quality of crowdsourced data contributed by expert and non-experts," *PLoS ONE*, vol. 8, no. 7, p. e69958, July 2013.

[15] H. Hutt, R. Everson, M. Grant, J. Love, and G. Littlejohn, "How clumpy is my image? Evaluating crowdsourced annotation tasks," in *Proc. UK Workshops on Computational Intelligence (UKCI)*, 2013, pp. 136–143.

[16] C. Dumitru, S. Cui, G. Schwarz, and M. Datcu, "Information content of very-high-resolution SAR images: Semantics, geospatial context, and ontologies," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 4, pp. 1635–1650, April 2015.

[17] S. Cui, "Spatial and temporal SAR image information mining," Ph.D. dissertation, University of Siegen, 2014.

[18] J. Singh, "Spatial content understanding of very high resolution synthetic aperture radar images," Ph.D. dissertation, University of Siegen, 2014.

[19] Y. Yang and S. Newsam, "Bag-of-Visual-Words and spatial extensions for land-use classification," in *Proc. ACM International Conference on Advances in Geographic Information Systems (GIS)*, 2010, pp. 270–279.

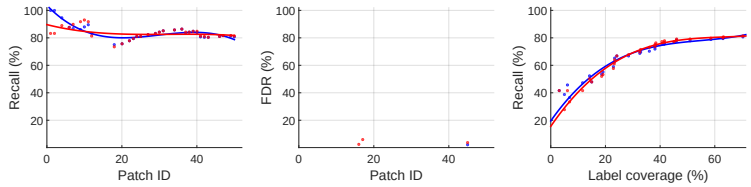
[20] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 173–176, January 2011.

[21] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

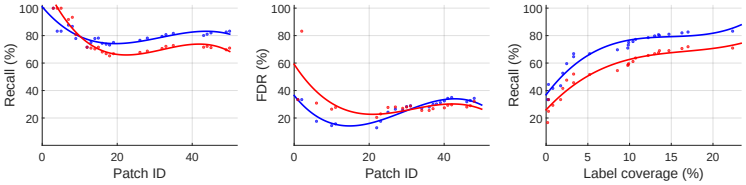
[22] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, September 2010.

[23] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, August 1996.

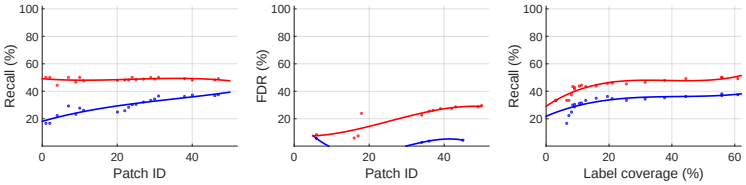
[24] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.
[25] M. Sabou, K. Bontcheva, and A. Scharl, "Crowdsourcing research opportunities: Lessons from natural language processing," in *Proc. International Conference on Knowledge Management and Knowledge Technologies*, 2012, pp. 1–8.
[26] L. Fortson and S. Lynn, "Talking in the zooniverse: A collaborative tool for citizen scientists," in *Proc. International Conference on Collaboration Technologies and Systems (CTS)*, 2014, pp. 1–2.
[27] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.
[28] C. Dumitru and M. Datcu, "Information content of very high resolution SAR images: Study of feature extraction and imaging parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 8, pp. 4591–4610, August 2013.
[29] A. Hanbury, "A Survey of Methods for Image Annotation," *Journal of Visual Languages & Computing*, vol. 19, no. 5, pp. 617–627, October 2008.



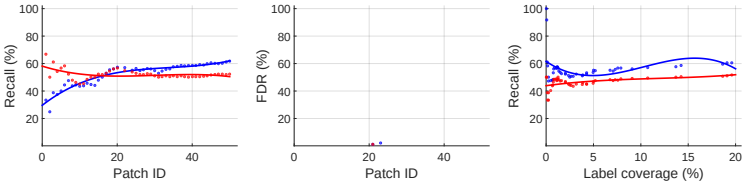
(a) Agricultural field



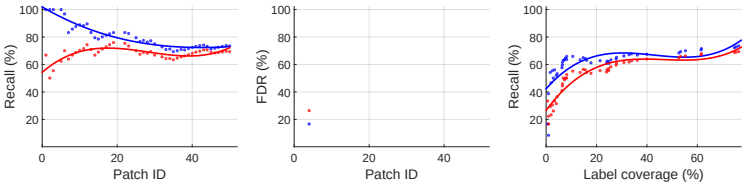
(b) Buildings



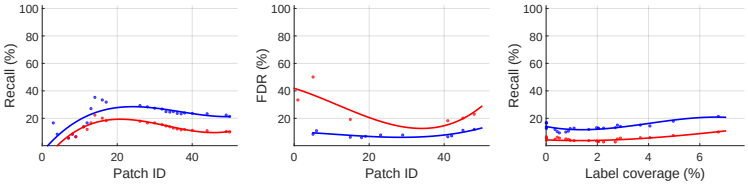
(c) Crop



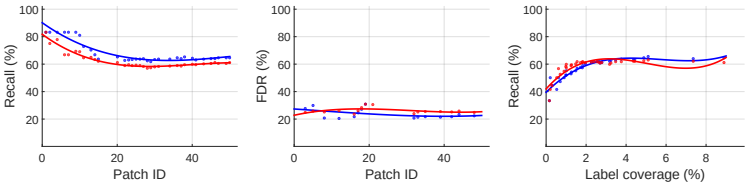
(d) Grass



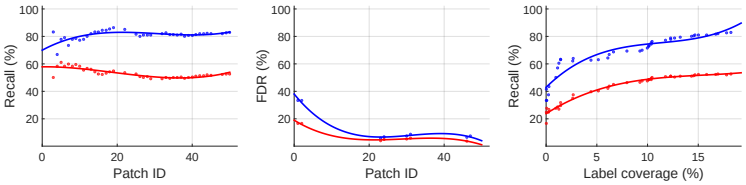
(e) House



(f) Parking lot



(g) Road



(h) Street

Fig. 10. User performance vs. the Patches in order of appearance, and label coverage for various labels. The red points in the graphs refer to UX_B_{MI}, whereas the blue points refer to UX_B_C.